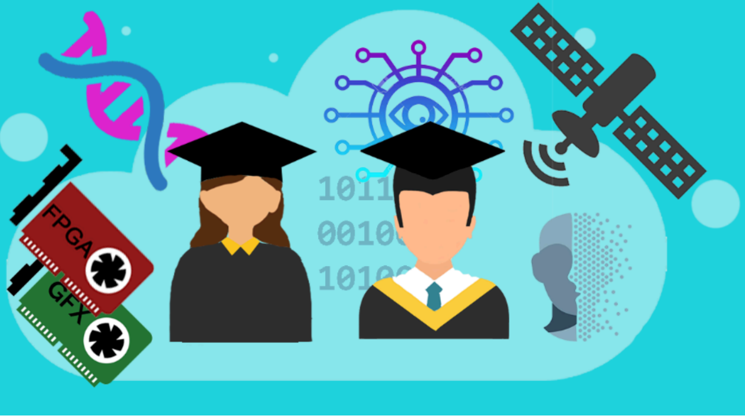
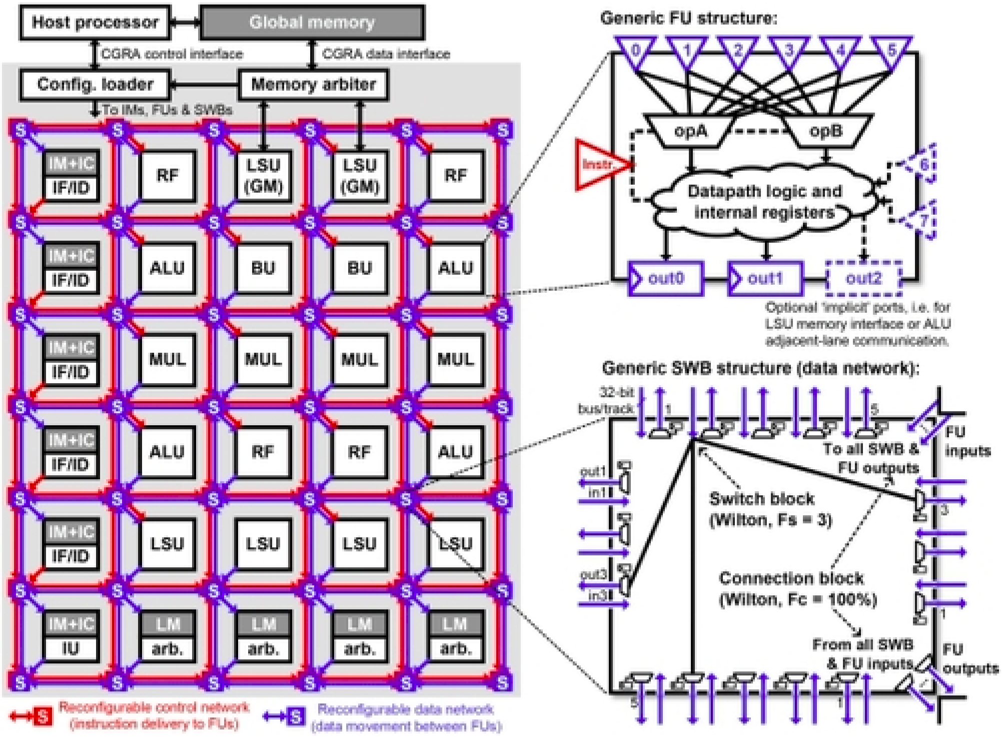
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Modeling ML applications on R-Blocks CGRA

Coarse-grained reconfigurable architectures (CGRAs) are programmable hardware accelerators composed of memory and processing tiles arranged in a 2D matrix. Specifically, R-Blocks [1] is a flexible ultra-low-power (ULP) CGRA developed by TuE, supported by a hardware/software tool-flow for code generation that includes VLIW and SIMD-like operations. R-Blocks facilitates the seamless integration of new tile designs within the CGRA framework, leading to reduced design and testing times, making it well-suited for optimizing signal processing and machine learning (ML) applications.


Host processor
Global memory
CGRA control interface
CGRA data interface
Config. loader
Memory arbiter
To IMs, FUs & SWBs
IM+IC
IF/ID
IU
RF
LSU (GM)
ALU
BU
MUL
LSU
LM arb.
S
Reconfigurable control network (instruction delivery to FUs)
Reconfigurable data network (data movement between FUs)
Generic FU structure:
0
1
2
3
4
5
6
7
opA
opB
Instr.
Datapath logic and internal registers
out0
out1
out2
Optional 'implicit' ports, i.e. for LSU memory interface or ALU adjacent-lane communication.
Generic SWB structure (data network):
32-bit bus/track
To all SWB & FU outputs
FU inputs
out1
in1
out3
in3
Switch block (Wilton, Fs = 3)
Connection block (Wilton, Fc = 100%)
From all SWB & FU inputs
FU outputs

Accurate application mapping and power-performance estimation are critical in digital design, as they minimize exploration, design, and testing time. Zigzag is a design space exploration (DSE) framework for hardware accelerator architecture and mapping, with a focus on ML applications [2]. By modeling simple TPU-like, in-memory computing (IMC) [3], or multi-processor [4] accelerators and importing neural network (NN) models in ONNX format, Zigzag generates optimal mappings and estimates power and performance metrics for each layer. However, Zigzag cannot map to the R-Blocks CGRA, as there is no corresponding hardware model that supports the architecture and tile flexibility of R-Blocks.

The main target of the thesis is to use the exploration capabilities of Zigzag and produce optimal mappings and accurate performance estimations for ML workloads by importing a model of the

R-Blocks CGRA. Version of the CGRA will be implemented in ASIC tools or FPGA hardware to evaluate the accuracy and efficiency of the extended framework.

The diploma thesis focuses on the following areas (but is not limited to):

- Modeling the energy and performance of R-Blocks CGRA on ML workloads.
- Exploration of efficient mappings of Neural Network Operations on the R-Blocks CGRA.
- Automatic C code generation from mapping output
- Validation of the Energy-Performance model on FPGA/ASIC toolflow (Global Foundries 22nm).

RELATED MATERIAL:

[1] Barry de Bruin, Kanishkan Vadivel, Mark Wijtvliet, Pekka Jääskeläinen, and Henk Corporaal. 2024. R-Blocks: an Energy-Efficient, Flexible, and Programmable CGRA. ACM Trans. Reconfigurable Technol. Syst. 17, 2, Article 34 (June 2024), 34 pages. https://doi.org/10.1145/3656642

[2] L. Mei, P. Houshmand, V. Jain, S. Giraldo and M. Verhelst, "ZigZag: Enlarging Joint Architecture-Mapping Design Space Exploration for DNN Accelerators," in *IEEE Transactions on Computers*, vol. 70, no. 8, pp. 1160-1174, 1 Aug. 2021, doi: 10.1109/TC.2021.3059962

[3] J. Sun, P. Houshmand and M. Verhelst, "Analog or Digital In-Memory Computing? Benchmarking through Quantitative Modeling," Proceedings of the IEEE/ACM Internatoinal Conference On Computer Aided Design (ICCAD), October 2023

[4] A. Symons, L. Mei, S. Colleman, P. Houshmand, S. Karl and M. Verhelst, “Towards Heterogeneous Multi-core Accelerators Exploiting Fine-grained Scheduling of Layer-Fused Deep Neural Networks”, *arXiv e-prints*, 2022. doi:10.48550/arXiv.2212.10612.

PREREQUISITES:

- Python, Bash, Digital Design/FPGA
- Familiarity with Machine Learning/Neural Networks

CONTACT INFORMATION:

- Panagiotis Chaidos, Junior Researcher, NTUA: (pchaidos@microlab.ntua.gr)
- Alexis Maras Ph.D Student, NTUA: (amaras@microlab.ntua.gr)
- Georgios Alexandris, Ph.D Student, NTUA: (galexandris@microlab.ntua.gr)
- Professor Dimitrios Soudris, NTUA: (dsoudris@microlab.ntua.gr)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)